A High-Speed Algorithm for Genome-Wide Association Studies on Multi-GPU

Poster Summary

Yen-Chen Chen
Institute of Applied Mathematical Sciences, National Taiwan University
Taipei, Taiwan
yanjen@ntu.edu.tw

Su-Yun Huang (Advisor)
Academic Sinica
Taipei, Taiwan

Shih-Jen Tsai (Advisor)
Taipei Veterans General Hospital
Taipei, Taiwan

Albert C. Yang (Advisor)
Harvard Medical School
Boston, Massachusetts

Weichung Wang (Advisor)
Institute of Applied Mathematical Sciences, National Taiwan University
Taipei, Taiwan

1 INTRODUCTION

We develop a high speed algorithm for Genome-Wide Association Studies which can run parallelly on multiple GPUs.

1.1 Genome-Wide Association Studies

Genome-wide association studies (GWAS)[1] have been a powerful tool for investigating the genetic architecture of human diseases. Scientists have found various of traits depending SNPs in the past decades by GWAS.
Single nucleotide polymorphisms (SNPs) could provide useful information regarding genetic variance of a population. Compare to the full 3 billion DNA series, SNPs have size of order 10^5, which is easier for computation. The problem we solve is GWAS for SNPs with brain parameters as traits.

1.2 Brain Parameters

Although scientists have studied human genes for decades, how these genetic markers are associated with brain function remain largely unknown. One critical reason is that we cannot determine a single representative trait for the whole brain. Different part of the brain is in charge of different part of our body, including visualization, language, and more. The part in a brain that actually controls our body is the gray matter. In a 3D brain fMRI, we can find 55749 voxels of gray matter, which is what we're curious of.

1.3 Detailed description of our data

The study constituted a continuing effort of the Healthy Aging Project[7, 8], and received approval from the Institutional Review Board of Taipei Veterans General Hospital. Each participant gave written informed consent and was evaluated by a trained research assistant using the Mini-International Neuropsychiatric Interview to exclude the presence of Axis I psychiatric disorders[5]. All participants were assessed for cognitive function, using the Mini-Mental State Examination (MMSE)[2] and the Wechsler Digit Span Task[6]. Older participants were further assessed using the Clinical Dementia Rating Scale (CDR)[4] to exclude dementia (CDR > 0). Overall exclusion criteria for all participants were as follows: (a) presence of dementia; (b) presence of Axis I psychiatric disorders, such as schizophrenia, bipolar disorders, or unipolar depression; and (c) a history of neurological conditions, such as head injury, stroke, or Parkinsons disease.

2 DIFFICULTIES

GWAS requests high accuracy for p value. And there is no robust rule for how small p value should be to count as important factor. People usually choose p value less than 10^{-6} or 10^{-7}, which requires accurate calculation for p value. To calculate p value, we will have to deal with the unbounded numerical integration:

$$p = \int_{-\infty}^{t} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \times \frac{1}{\sqrt{\nu\pi}} \times \frac{dt}{(1+\frac{t^2}{\nu})^{\frac{\nu+1}{2}}}$$

which is not ideal for GPU implementation. In order to implement this numerical integration in GPU, we have to come up with another algorithm.

Why compute p value ? Most statistical studies with hypothesis testing picks a standard t value as a threshold to reject the false hypothesis. In fact, previous study in GPU parallelized GWAS[3] didn't include p value computation. However, we decide to compute p value directly with the following reasons.

(1) Since we have 36 billion sets of t-tests, we have no robust theroy for what t value threshold to choose.
(2) There's no existing t table for extreme small p value with large degrees of freedom.
(3) Due to the loss of SNP data for some patients, the degrees of freedom is not a fixed value.

3 OBJECTIVES

We want to know how all 653291 SNPs affect 55749 voxels of gray matter brain function parameters.
To solve this GWAS problem, we choose to use Single Locus Analysis, which is to solve simple linear model for each SNP to brain voxel pair.

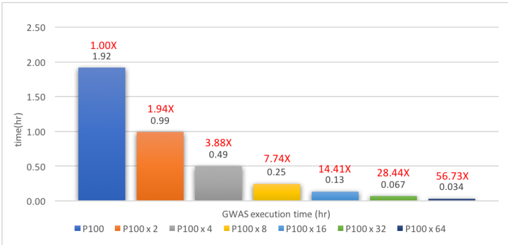

Figure 1: GWAS P100 multi-GPU scalability

For each SNP to brain voxel pair, we have a linear model:

$$Y = X\beta + \epsilon$$

Where β is some parameter vector, and ϵ is error term. We will first solve a linear regression:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

and calculate a p-value for each of the t-tests:

$$t = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 (X^\top X)^{-1}_{ii}}}$$

$$p = \int_{-\infty}^{t} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \times \frac{1}{\sqrt{\nu\pi}} \times \frac{dt}{(1+\frac{t^2}{\nu})^{\frac{\nu+1}{2}}}$$

X: Data matrix, Y: Vector of brain responses
σ: Standard deviation, ν: Degrees of freedom, Γ: Gamma function
Parallelization
However, all linear systems are independent and thus this GWAS problem is perfect for parallelization.

4 REFORMATION

In order to implement p value calculation in GPU, we come up with a new reformation.

$$p = C \int_{-\frac{\pi}{2}}^{\arctan(\frac{t}{\sqrt{\nu}})} \cos^{\nu-1}\theta d\theta, \quad C = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \times \frac{1}{\sqrt{\pi}}$$

Where C is a constant shared by all integrations.
By implying simple calculus, the reformation of the p value integration make it much simpler and efficient for GPU implementation. This allows us to get moch more precise p value with the same computational time.
Note that the upper integration is **bounded**.

5 RESULTS

5.1 Runtime Comparison

Table 1 is our runtime table. A simple C++ code with the same computating method would take 1.5 months to finish the whole calculation. We develop a cuda code that runs a lot faster than CPU code. The computation time can be reduced to 2 minutes with 64 nvidia P100 GPUs.
Note that the runtime are with reformed p value integration.

[1]C++ runtime is approximated.

	GPU	nodes	time(hour)
C++	-	1	1065.80[1]
cuda	K40	1	10.55
cuda	K40	2	5.26
cuda	P100	1	1.92
cuda	P100	2	0.99
cuda	P100	4	0.49
cuda	P100	8	0.25
cuda	P100	16	0.13
cuda	P100	32	0.067
cuda	P100	64	0.034

Table 1: GWAS runtime comparison

The algorithm also has high scalability as shown in Figure 1.

5.2 Numerical Results

We can sum up all $-\log p$ values and project back to a standard brain to find out brain sections easily affected by SNPs.

6 CONCLUSION

We develop a fast multi-GPU algorithm and cuda code for p value calculation. This code can deal with large GWAS problems which have never been done before.

ACKNOWLEDGMENTS

We are grateful to Professor Kengo Nakajima for helpful discussions and the Information Technology Center of The University of Tokyo for the usage of the Reedbush-H system.

REFERENCES

[1] William S Bush and Jason H Moore. 2012. Genome-wide association studies. *PLoS Computational Biology* 8, 12 (2012), e1002822.
[2] Marshal F Folstein, Susan E Folstein, Paul R McHugh, et al. 1975. Mini-mental state. *A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res* 12, 3 (1975), 189–198.
[3] Jorge González-Domínguez, Bertil Schmidt, Jan Christian Kässens, and Lars Wienbrandt. 2014. Hybrid CPU/GPU acceleration of detection of 2-SNP epistatic interactions in GWAS. In *European Conference on Parallel Processing*. Springer, 680–691.
[4] Charles P Hughes, Leonard Berg, Warren L Danziger, Lawrence A Coben, and RL1982 Martin. 1982. A new clinical scale for the staging of dementia. *The British journal of psychiatry* 140, 6 (1982), 566–572.
[5] D Sheehan, Yves Lecrubier, K Harnett Sheehan, K Sheehan, P Amorim, J Janavs, E Weiller, T Hergueta, R Baker, and G Dunbar. 1998. Diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatr* 59 (1998), 22–33.
[6] D Wechsler. 1997. WAIS-III: Wechsler adult intelligence scale, administration and scoring manual. San Antonio, Texas: Psychological Corporation. (1997).
[7] Albert C Yang, Chu-Chung Huang, Mu-En Liu, Yin-Jay Liou, Chen-Jee Hong, Men-Tzung Lo, Norden E Huang, Chung-Kang Peng, Ching-Po Lin, and Shih-Jen Tsai. 2014. The APOE 4 allele affects complexity and functional connectivity of resting brain activity in healthy adults. *Human brain mapping* 35, 7 (2014), 3238–3248.
[8] Albert C Yang, Chu-Chung Huang, Heng-Liang Yeh, Mu-En Liu, Chen-Jee Hong, Pei-Chi Tu, Jin-Fan Chen, Norden E Huang, Chung-Kang Peng, Ching-Po Lin, et al. 2013. Complexity of spontaneous BOLD activity in default mode network is correlated with cognitive function in normal male elderly: a multiscale entropy analysis. *Neurobiology of aging* 34, 2 (2013), 428–438.